FINAL PROGRESS REPORT

Identification of Patients with Low Life Expectancy

Principal Investigator
Alexander Turchin, MD, MS[1,2]

Team Members
Leonid Perlovsky, PhD[3]
Wendong Ge, PhD[1,2]
Hee Yeun Kim, PhD[1,2]
Shervin Malmasi, PhD[1,2]
Fritha Morrison, MPH[1]
Alex Solomonoff, PhD[1,2]
Maria Shubina, ScD[1]

Organization
Brigham and Women's Hospital

Project Dates: 08/01/2015 – 05/31/2019

Federal Project Officer: Parivash Nourjah

Acknowledgement of Support: This study was funded by the United States Agency for HealthCare Research and Quality grant # 1-R01-HS024090.

Grant Award Number: 1-R01-HS024090

[1]Division of Endocrinology, Brigham and Women's Hospital, Boston, MA
[2]Harvard Medical School, Boston, MA
[3]LP Information Technology, Brookline, MA

Structured Abstract

Purpose: To test whether a novel machine learning algorithm Dynamic Logic in conjunction with natural language processing can achieve higher prediction accuracy for the risk of death over the next 12 months compared to the benchmark statistical and machine learning algorithms.

Scope: We tested Dynamic Logic and benchmark algorithms on a study population of 630,000 patients ≥ 40 years of age treated in an integrated healthcare delivery system between 2000 and 2014.

Methods: A unique patient-year, starting on January 1st and ending on December 31st of a given year served as the unit of analysis. A single patient could contribute multiple units of analysis. Data were obtained from electronic medical records. Study dataset included variables characterizing patient's demographics, diagnoses, procedures, vitals and laboratory tests, as well as data obtained from narrative provider notes. Benchmark algorithms included logistic regression, support vector machines and neural networks.

Results: Dynamic Logic had a small but consistent advantage in estimating the probability of death over the benchmark methods. Data normalization and algorithm optimization methods were significant contributors to algorithm accuracy. Information from text significantly increased prediction accuracy.

Key Words: predictive modeling, machine learning, mortality, electronic medical records

Purpose

This research project was funded to evaluate whether using artificial intelligence and natural language processing could improve identification of patients at high risk of death through the following project aims:

Specific Aim 1
To determine whether a combination of artificial intelligence technology Dynamic Logic and natural language processing improves accuracy of identification of patients with low life expectancy (likely to die over the next 12 months) compared to the currently used methods.

Specific Aim 2
To determine whether the low life expectancy model developed on the general patient population is equally accurate in patients with chronic conditions on the example of patients with a) diabetes, b) hypertension and c) osteoporosis.

Specific Aim 3
To develop open source software that will integrate Dynamic Logic and natural language processing to allow users to easily assess patients' life expectancy based on a combination of structured and narrative EMR data.

Scope (Background, Context, Settings, Participants, Incidence, Prevalence)

It is increasingly recognized that optimal treatment is not the same for every patient – it depends on the individual patient's circumstances. One important factor that determines the optimal clinical management is the patient's life expectancy. Individual's life expectancy helps establish the optimal clinical management because it determines the temporal horizon within which medical decisions have to operate. The effect of many chronic disease interventions may not be seen for years, or sometimes even decades. Therefore, while adding an extra diabetes medication may save a 40-year-old individual from going blind or developing kidney failure 20 years later, it will not bring any benefits to an 80-year old with a metastatic malignancy who is expected to live only a few months.

Consequently, it is critical that when we measure quality of care delivered by providers, suggest treatment options to clinicians through clinical decision support or compare different treatment strategies, we take into account the patient's life expectancy. However, currently there are no methods available that can do this with sufficient accuracy. Currently used methods typically only reach c-statistic in the 0.7 – 0.8 range which leads to large trade-offs in either sensitivity or specificity.

Most commonly used techniques to assess a patient's mortality risk draw primarily on administrative data, and sometimes on other structured data fields in electronic medical records. This approach leaves out a large amount of information that is only available in narrative documents such as provider notes, radiology and pathology reports, etc. In this project we proposed to test the hypothesis that application of two novel technologies could leverage the information in narrative electronic documents to significantly improve the accuracy of identification of patients with low life expectancy.

The first of these technologies was Dynamic Logic, developed by Dr. Perlovsky who was a co-investigator on this project. Dynamic Logic allows to circumvent the challenge of combinatorial complexity that limits the number of variables and their combinations that can be considered as predictors of an outcome by most currently used analytical methods. Dynamic Logic makes use of a limited number of iterative approximations to reduce the complexity of a problem with multiple predictor variables from exponential to approximately linear. Utilization of Dynamic Logic will allow us to greatly increase the richness of the models for identification of patients with low life expectancy and ultimately improve their accuracy.

The second novel technology that was going to be tested in this proposal is natural language processing (NLP). As electronic medical records (EMRs) grow increasingly prevalent, narrative documents become available for analysis. Previous research has indicated that information such as the patient's functional status that is usually only found in narrative documents may be critical to improving accuracy of identifying frail patients at high mortality risk. Modern NLP techniques can effectively identify key concepts in medical text but until now analytical methods allowed consideration of only a few of pre-selected concepts in prediction models. In this project we aimed to test whether combining NLP with Dynamic Logic could allow us to expand the number of concepts from narrative text that could be included in the limited life expectancy prediction model, leading to an improvement in accuracy.

Methods

Study Design

The overall study design was a retrospective cohort analysis of the relationship between patient characteristics at baseline and death within 12 months in a large dataset of EMR data using several statistical and machine learning techniques. Accuracy of different statistical and machine learning techniques was compared.

Patient Population

This project was conducted at Partners HealthCare – an integrated healthcare delivery system in eastern Massachusetts that was founded by Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH). Partners HealthCare has been using advanced EMR systems (both in- and outpatient) since 2000. These EMR systems include records for medications, adverse reactions to medications, diagnoses / problems, procedures, family history, vital signs, laboratory data, patient demographics and narrative documents, such as provider notes. EMR data are available to Partners investigators through Research Patients Data Registry (RPDR). RPDR data also includes dates of death for all Partners HealthCare patients obtained from Social Security Death Master File.

The study analyses included patients 40 years and older who have been followed in a primary care practice affiliated with MGH or BWH for at least 12 months between 2000 and 2014 as evidenced by at least 2 notes during that period of time. Based on previous experience we estimate that 85% of healthcare utilization by patients treated in Partners HealthCare primary care practices takes place at Partners HealthCare facilities, allowing us to capture maximum amount of clinical information that could be relevant for prediction of low life expectancy. Based on these criteria, the study included 630,000 patients. The study patient population was randomly divided into the *training dataset* that contained 80% of the study patient population and the held-out *validation dataset* that contained 20% of the study patient population.

Variable Definitions

As the risk of death for a given patient may change over their lifetime as their age, comorbidities and other characteristics change, we evaluated the risk of death at a particular point in time. Specifically, we evaluated the risk of death once a year, on January 1st. A unique patient-year that began on January 1st of a given year (the index date) and ended on December 31st of the same year therefore served as the unit of analysis. A single patient could contribute multiple units of analysis to the study. A particular patient-year was only included in the analysis if they had at least one primary care practice encounter prior to the index date (and therefore were likely to have reliable baseline information). The training dataset included 1.6 million patient-years and the validation dataset included 400,000 patient-years.

Table 1. Study Variables.

Category	Variable	Description
Demographics	Sex	
	Age	At the index date
	Race	
	Education level	
Diagnoses[1]	DxEver	Binary variable that is set to "1" if the patient ever had this ICD code recorded prior to the index date and "0" otherwise.
	DxCount	Numerical variable that indicated the number of times this ICD code was recorded in the 12 months prior to the index date.
Procedures	ProcedureEver	Binary variable that is set to "1" if the patient ever had this CPT code recorded prior to the index date and "0" otherwise.
Labs[2]	LabEver	Binary variable that is set to "1" if the patient ever had this CPT code recorded prior to the index date and "0" otherwise.
	LabSlope	Numerical variable indicating the slope of the line fitted through the laboratory test results over 12 months prior to the index date.
	LabSD	Numerical variable indicating the standard deviation of the laboratory test results over 12 months prior to the index date.
Meds[3]	MedEver	Binary variable that is set to "1" if the patient ever had this medication recorded.
	MedLastYear	Binary variable that is set to "1" if the patient had this medication recorded within 12 months prior to the index date.
	MedMaxDose	Numerical variable that is set to the maximum dose of this medication the patient had recorded within 12 months prior to the index date.
Vitals	Pulse	Patient's last recorded heart rate before the index date
	SBP	Patient's last recorded systolic blood pressure before the index date
	DBP	Patient's last recorded diastolic blood pressure before the index date
	BMI	Patient's last recorded body mass index before the index date
Outcome	Deceased	Binary variable that is set to "1" if the patient died within 12 months from the index date.

[1]One set of *diagnosis*-related variables was generated for each ICD code
[2]One set of *lab*-related variables was generated for each laboratory test
[3]One set of *medication*-related variables was generated for each unique active ingredient-route group combination

Each patient-year record included information on the patient's demographics, diagnoses, procedures, vital signs and laboratory tests (Table 1). The variables were designed to reflect both static characteristics (e.g. does the patient have diagnosis X) as well as their dynamic aspects. For example, we included a variable representing the number of times a particular ICD code was recorded in the previous 12 months because it may reflect the condition's acuity. We also included a variable representing the slope of the line fitted to the laboratory test measurements may reflect stability vs. deterioration of a particular test. Laboratory measurements were log-transformed because we empirically found that this was a more effective representation of their large dynamic range.

Benchmark Methods

We utilized several benchmark methods analytical for prediction of risk of death. The first one was logistic regression. It was implemented using scikit-learn python library with 12-fold cross-validation. We also utilized two machine-learning benchmark techniques: support vector machines (SVMs) and neural networks. Both of these were also implemented using scikit-learn python library. We have tested neural networks with one, two and three hidden layers.

Dynamic Logic

Dynamic Logic is a machine-learning technique invented by our co-investigator, Dr. Leonid Perlovsky. The Dynamic Logic process begins with a vague-fuzzy state (or model) and within few iterations it converges to a near exact (crisp) state fitting the data. Whereas multiple hypothesis testing routinely requires sorting through a large number of combinations of data, the Dynamic Logic process avoids combinations; within few iterations it "jumps" from a vague state to a solution. Dynamic Logic starts with vague-fuzzy assignment of each data point to all groups (in other words, each data point has a "probability" to be in each group). Regression equations are then computed for each group. On the next iteration assignment "probabilities" are modified so that the average of all regression predictions improves a bit. After a small number of iterations (typically several dozen or fewer) the best possible choice of groups and most accurate regression equations are attained. The mathematical innovation of Dynamic Logic is achieved in developing the mathematical description of the process, which iteratively improves both assignments of data to the groups and regression equations; this joint process is the fundamental innovation of Dynamic Logic.

In our implementation of Dynamic Logic count variables (e.g. DxCount) were modeled as Poisson random variables and real-valued variables (e.g. LabSlope or vitals-related variables) were modeled using a Gaussian distribution.

We have tested several approaches to the optimization component of the Dynamic Logic algorithm: a) expectation-maximization (EM) maximum likelihood; b) discriminative training; and c) optimal area-under-the-curve (AUC) training. We additionally tested several approaches to initialization, including vague and wide-vague initialization. We also tested a variety of approaches of increasing complexity to regularization of the real-valued (Gaussian) variables. We also tested using a *dustbin* cluster - a cluster that is supposed to absorb all of the records that aren't modeled well by the other clusters (i.e. a featureless cluster with large variation).

Computational Text Analysis

We evaluated two off-the-shelf packages for natural language processing: MetaMap and cTAKES. We found that both had very slow performance – an unacceptable limitation for our very large training dataset. We also found that they had relatively low accuracy in mapping text to UMLS concepts, negating the possible advantage of synonym identification. Finally, cTAKES also required real-time UMLS license-checking, introducing additional technical challenges. Based on this, we ultimately chose to implement our own text analysis algorithms. These algorithms accomplished the following tasks: a) identification of document boundaries and linking individual documents to patient-years (units of analysis in our study); b) identification and removal of text that does not carry clinical information (e.g. HTML tags); c) tokenization (identification of word boundaries); d) calculating the number of occurrences for each unique word in each document as well as in the overall dataset; e) excluding rarely found unique words (likely to be misspellings) and f) TF-IDF normalization. TF normalization involves dividing word counts for each unique word by the total number of words in that document and IDF normalization involves scale each word by a function of what fraction of documents it occurs in.

Evaluation

C-statistic (area under the ROC curve) was used to determine the prediction accuracy of the Dynamic Logic algorithm and to compare it to the benchmark methods. C-statistic for both Dynamic Logic algorithm and the benchmark methods was calculated on the held-out validation dataset after algorithm parameters were optimized using cross-validation on the training dataset.

Infrastructure

The project was performed on a large Linux-based cluster. Hyperparameter optimization using grid search cross-validation is very computationally intensive, particularly on a large training dataset used in this project. We therefore developed software for automated parceling of the overall task into subcomponents and batch submission of these subcomponents to the cluster.

Results

Specific Aim 1

To determine whether a combination of artificial intelligence technology Dynamic Logic and natural language processing improves accuracy of identification of patients with low life expectancy (likely to die over the next 12 months) compared to the currently used methods.

Including greater amount of information generally improved algorithm performance (although usually required a significant amount of effort for optimization of analysis of every individual variable category). For example, adding procedure information to diagnoses improved performance slightly; Dynamic Logic exceeded performance of Logistic Regression (Table 2).

Table 2.
Impact of Adding Procedures to Diagnoses Information on Death Risk Prediction Accuracy

System	Diagnoses Included	Procedures Included	AUC
Logistic Regression	✓	-	0.8900
Dynamic Logic	✓	-	0.8918
Logistic Regression	✓	✓	0.9160
Dynamic Logic	✓	✓	0.9200

Dynamic Logic algorithm that utilized discriminative training performed slightly better than the one using EM maximum likelihood, when combined with wide-vague initialization (Table 3).

Table 3. Effect of Training and Initialization Approach on Dynamic Logic Algorithm Accuracy

System	Training	Initialization	AUC
Logistic Regression	-	-	0.9160
Dynamic Logic	EM-ML	standard	0.9158
	Discriminative	wide-vague	0.9186

The Dynamic Logic algorithm had a number of parameters that had to be optimized through hyperparameter optimization. Due to the large number of parameters (in part due to the complicated regularization for Gaussian variables), grid search hyperparameter optimization that performs exhaustive testing of all parameter combinations was not feasible. Therefore an algorithm had to be used to select the parameter combinations to be tested. In addition to the baseline hill-climbing algorithm for selection of parameter combinations, we also tested Best Average Neighbor Score (BANS) algorithm with both 30 and 60 points on the parameter grid being tested simultaneously. This testing was performed separately for the entire study population (Table 4) and for the subpopulation of patients aged ≥ 65 (Table 5). As expected, predictive performance was lower for the subpopulation of patients aged ≥ 65. However, this was likely closer to the true algorithm performance as predictive accuracy on the entire study population (age ≥ 40) may have been inflated due to the low frequency of death among younger individuals.

Table 4. Comparison of Dynamic Logic Optimization Algorithms on the Entire Study Population

System	Optimization	AUC
Logistic Regression	N/A	0.92617
Dynamic Logic	Baseline	0.92739
Dynamic Logic	BANS – 30 points	0.92757
Dynamic Logic	BANS – 60 points	0.92935

Table 5. Comparison of Dynamic Logic Optimization Algorithms on Patients Aged ≥ 65

System	Optimization	AUC
Logistic Regression	N/A	0.87075
Dynamic Logic	Baseline	0.87646
Dynamic Logic	BANS – 30 points	0.87722

Dynamic Logic algorithm that incorporated hyperparameter optimization in this way was compared to other machine learning algorithms, including SVM and neural networks, on both the entire study population (Table 6) and the subpopulation of patients aged ≥ 65 (Table 7).

Table 6. Dynamic Logic vs. Benchmark Algorithms on the Entire Study Population

System	AUC
Logistic Regression	0.92617
SVM	0.92752
Dynamic Logic	0.92935

Table 7. Dynamic Logic vs. Benchmark Algorithms on Patients Aged ≥ 65

System	AUC
Logistic Regression	0.87075
SVM	0.87201
Neural Network: 1 hidden layer	0.8735
Neural Network: 2 hidden layers	0.8740
Neural Network: 3 hidden layers	0.8745
Dynamic Logic	0.87722

Including word count variables into the Dynamic Logic model significantly increased its performance: AUC for the subpopulation of patients aged ≥ 65 increased to 0.9469. The words that had the greatest impact on the probability of death included a number of terms with direct implications for severity of illness and / or overall prognosis: *hospice*, *normal*, *metastatic*, *palliative*, and *admitted*.

Specific Aim 2
To determine whether the low life expectancy model developed on the general patient population is equally accurate in patients with chronic conditions on the example of patients with a) diabetes, b) hypertension and c) osteoporosis.

We were unable to accomplish Specific Aim 2 because the sample size was insufficient. Patients with diabetes, hypertension and osteoporosis constituted 5 – 15% of the overall data sample. That resulted in the training dataset of < 240,000 patient-years. Given that the dataset included 239,000 features (variables), it was not possible to effectively train machine learning models that had 1:1 ratio of features to records.

Specific Aim 3
To develop open source software that will integrate Dynamic Logic and natural language processing to allow users to easily assess patients' life expectancy based on a combination of structured and narrative EMR data.

The software that integrates Dynamic Logic and natural language processing to assess patients' life expectancy is available upon request to the PI Alexander Turchin who can be contacted at aturchin@bwh.harvard.edu.

Conclusions

Based on these results, it can be concluded that Dynamic Logic had a small but consistent advantage in estimating the probability of death over the next 12 months over the benchmark methods, including logistic regression, support vector machines and neural networks. Data normalization and algorithm optimization methods were significant contributors to algorithm accuracy – more important than the actual algorithm chosen. Including information from text significantly increased prediction accuracy, and has a potential for being a critical component of future predictive algorithms for risk of death.

List of Publications and Products

None.